

BINOMIAL REGRESSION IN GLIM: ESTIMATING RISK RATIOS AND RISK DIFFERENCES¹

SHOLOM WACHOLDER

Wacholder, S. (Dept. of Epidemiology and Biostatistics, McGill U., Montreal, PQ, H3A 1A2, Canada). Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 1986;123:174-84.

Although an estimate of the odds ratio adjusted for other covariates can be obtained by logistic regression, until now there has been no simple way to estimate other interesting parameters such as the risk ratio and risk difference multivariately for prospective binomial data. These parameters can be estimated in the generalized linear model framework by choosing different link functions or transformations of binomial or binary data. Macros for use with the program GLIM provide a simple method to compute parameters other than the odds ratio while adjusting for confounding factors. A data set presented previously is used as an example.

biometry; regression analysis; statistics

One of the many reasons for the popularity of logistic regression in epidemiologic applications is the flexibility it allows in the choice of covariates to be included in the model. However, estimators of parameters other than the odds ratio for prospective binomial data, such as the risk ratio and the risk difference, have been available only in simple situations (1, 2). Noteworthy is the Mantel-Haenszel (3) type risk ratio estimator for stratified data, whose variance is discussed by Breslow (2). In this paper, I show how risk ratio and risk difference parameters can be related to regression coefficients for fitting binomial data by assuming a functional relationship be-

tween disease probabilities and a linear combination of the covariates. This enables multivariate estimation of risk ratio or risk difference parameters while controlling for confounding and considering interaction. Using macros (sets of commands approximately analogous to subroutines) for the program GLIM (4) is a convenient way to obtain maximum likelihood estimates of these parameters when covariates are continuous, categorical, or both. More generally, when any monotone function of probabilities is assumed to be a linear function of the covariates, extension of these principles allows estimation of the parameters.

MODELS

In these regressions, the dependent variables are I observed proportions, each based on $N_i \geq 1$ independent observations. Each proportion has unknown probability of success π_i , where π_i is functionally related to the linear predictor, denoted by %LP in GLIM terminology, and defined when there are K covariates as

$\%LP = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K$,
the sum of the products of unknown regres-

Received for publication December 27, 1984, and in final form April 30, 1985.

¹ From the Department of Epidemiology and Biostatistics, McGill University, Montreal, PQ, Canada.

Reprint requests to Dr. Sholom Wacholder, McGill University, Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, Montreal, PQ, H3A 1A2, Canada.

This research was partially supported by Grant A8752 from the National Science and Engineering Research Council of Canada.

The author thanks Dr. Ben Armstrong for his help with this work.

sion coefficients β_k and the values of the covariate for the i th proportion. A smooth monotone transformation, or *link function*, relates the covariate values for each proportion to its respective probability through the linear predictor. For logistic regression, the link function used is the logarithm of the odds of the probabilities; symbolically, $\log(\pi/(1 - \pi)) = \%LP$. Inversion gives $\pi = \exp(\%LP)/(1 + \exp(\%LP))$. For risk difference models, the identity link function $\pi = \%LP$ is used; for the risk ratio, the link is $\log(\pi) = \%LP$, equivalent to $\pi = \exp(\%LP)$. Other monotonic link functions can be used. For example, the log-complement link $\log(1 - \pi) = \%LP$, described by Weinberg (5), models the log of the probability of no disease, or health, which has interesting applications for studying synergy. Probit, arc-sin, and complementary log-log transformations are discussed in the GLIM manual (4). Storer et al. (6) present macros for fitting the odds transformation.

In each of these models, the regression coefficient β_k represents the difference in the probability, transformed by the appropriate link function, associated with a unit change in the value of the covariate X_k when the other $K - 1$ covariates remain constant. For the logistic link, β_k is the difference between the logarithms of the odds, and, therefore, $\exp(\beta_k)$ is the ratio of the odds, associated with a change of one unit in X_k . Similarly, for the identity link, β_k is a risk difference, and for the logarithmic link, $\exp(\beta_k)$ is the risk ratio. For the log-probability of health link, β_k represents the logarithm of the "health ratio" or the ratio of the probabilities of no disease at values of X_k which differ by one unit.

The risk difference model is related to the additive relative risk model discussed by Thomas (7) and Storer et al. (6) for stratified data. The additive relative risk model essentially fits a model in the odds and presumes different baseline probabilities for each stratum.

Different links imply different models, and therefore, generally, different fits of

the data, even when the same covariates are included in the model. The link specifies the form of the relationship between the transformed probabilities and the covariate values. For example, a model with a logarithmic link and a single continuous covariate assumes a linear relationship between the covariate and the logarithm of the probability, while the identity link assumes that the linear relationship is between the covariate and the probability itself. For different links, therefore, qualitatively different dose-response relationships are fit.

Furthermore, when several factors are included in the model without an interaction, each link implies different constraints on the fitted probabilities. For example, if A and B are dichotomous risk factors, the contrast between $A = 2$ and $A = 1$ will be the same for both values of B ; but note that for each different link, the "effect" will be the difference in the *transformed* probability. Thus, with a logistic link, no interaction implies that the difference in the logits of the probabilities will be the same for $B = 1$ and $B = 2$, while for the identity link, no interaction implies equal differences of the untransformed probabilities. Generally, when A and B are independent risk factors, absence of interaction on one scale implies *presence* of interaction on other scales. Thus, for each link, there are different criteria for presence and absence of interaction, and use of different links allows for several assessments of synergy. The implications of different definitions of interaction are discussed elsewhere (8-10).

The fitted values for each link are identical, and the parameters (and their estimates) are functionally related only when the model is saturated; that is, when the number of fitted parameters equals the number of distinct covariate vectors. This occurs when the model fits main effects and all possible interactions of several polytomous variables; a special case of this is a single polytomous variable. Since a separate parameter will be available for each

cell with distinct covariate values, the fitted proportions will equal the observed proportions for all links.

ESTIMATION

In regressions with a binomial outcome variable, the assumption from classic regression of constant variance does not hold. A proportion based on N_i observations with mean π_i will have variance $\pi_i(1 - \pi_i)/N_i$. The asymptotic variance of the transformed proportion generally also depends on π_i . If the π_i 's were known, a weighted regression could be used, with each observation assigned a weight inversely proportional to its variance. The π_i 's, however, depend on the unknown regression coefficients.

GLIM solves this problem by using the iterative procedure described by McCullagh and Nelder (11). Weights are assigned to each proportion based on the estimates from the model of the π_i 's, which in turn are based on the estimates of the vector of β 's from the most recent iteration. Parameter estimates from successive iterations converge to the maximum likelihood estimate, regardless of the values of π_i used for the first iteration step (11), when the parameter space of $\beta = (\beta_0, \beta_1, \dots, \beta_K)$ is R^{K+1} , the full $K + 1$ dimensional space, allowing any possible vector of parameter values.

RESTRICTION OF THE PARAMETER SPACE

While the domain of all link functions is the 0-1 interval, the range of some, such as the logarithmic link, is not the set of real numbers. The inverses of these link functions may be undefined or lie outside the 0-1 interval for some values of %LP. For example, positive values of %LP for the logarithmic link and values of %LP outside the 0-1 interval for the identity link imply probabilities which do not make sense. This can cause problems in estimating risk ratios, health ratios, and risk differences when a continuous covariate or several polytomous covariates are included in the model. The odds ratio, based on the logit

link whose range is the set of all real numbers, does not share this problem. Thus, one advantage of using logistic regression is that there is no bound on the possible values of parameter estimates.

There are two possible reasons why impermissible values of %LP occur. Perhaps most likely is that the model itself is misspecified. The nonsense probabilities then serve as a warning of an inappropriate model. Even when the model is correct, however, random variation may result in fitted probabilities which are not between 0 and 1. This is prone to occur when some of the true probabilities are close to 1, to 0, or to either, for the logarithmic, log-complement, and identity links, respectively, especially when there is considerable spread in the true probabilities.

Disallowed values of %LP also imply impermissible values of the vector β . The parameter space for β , given a set of covariate vectors, is the subset of R^{K+1} which generates a vector of fitted probabilities whose components all lie between 0 and 1. When convergence occurs on the boundary of the restricted parameter space, the convergence will not generally be to the maximum likelihood estimate within the restricted parameter space. Thus, estimates of β leading to probabilities equal to 0 or 1 should be regarded as indications of model inadequacy rather than as reasonable parameter estimates.

The macros check for probabilities between 0 and 1 at each iteration. When a probability which is out of range is fitted, the linear predictor is revised arbitrarily to put it within range. The macros can be revised to do more extensive checking and warning or to delete the point which is out of range when an estimate on the boundary of the parameter space is suspected, such as when a fitted value very close to 0 or 1 is obtained.

ITERATION

There are four basic calculations which GLIM requires for each iteration in the

TABLE 1

Number of babies with birth weight less than the tenth percentile, by maternal alcohol consumption, smoking status, and social class*

Social class	Alcohol intake					
	Nonsmokers			Smokers		
	Heavy	Moderate	Light	Heavy	Moderate	Light
I and II	11/84	5/79	11/169	6/28	3/13	1/26
III	4/22	3/25	12/162	4/17	2/7	6/38
IV and V	0/14	1/18	12/91	7/19	2/18	8/70

* Reproduced with permission from Wright et al. (13).

TABLE 2

Deviances for different models and links

Model*	Parameter				
	OR	RR	RD	HR	dft
A. Grand mean only	33.2	33.2	33.2	33.2	17
B. ALC	24.0	24.0	24.0	24.0	15
C. SMO + SOC	22.9	22.8	22.9	22.9	14
D. ALC + SMO + SOC	13.8	13.6	14.9	15.1	12
E. ALC + SMO + SOC + ALC.SMO	12.0	12.0	11.6	11.6	10
F. ALC2 + SMO + SOC	14.2	14.2	15.2	15.3	13
G. ALC2 + SMO + SOC + ALC2.SMO	13.5	13.5	13.4	13.4	12
H. ALCC + SMO + SOC	14.3	14.3	15.9	16.2	13

* ALC, SMO, and SOC are categorical variables for maternal alcohol consumption, smoking status, and social class, respectively. ALC2 is a dichotomous variable with heavy drinkers in one category and moderate and light drinkers in the other category. ALCC is a continuous variable, assigning a 1 to light drinkers, 2 to moderate drinkers, and 3 to heavy drinkers.

† Degrees of freedom of the deviance.

maximum likelihood estimation for binomial data with a given link:

1. Calculation of %FV (the vector of fitted values), the product of N_i and the fitted proportion $\hat{\pi}_i$, which itself is based on the current value of the vector %LP. (As discussed above, the calculated fitted probabilities may be negative or greater than 1 for some observations. For example, if the %LP is positive for any observation when using the logarithmic link, the corresponding probability would be estimated to be greater than 1. %FV is checked to ensure that each component lies between 0 and N_i .)

2. Calculation of %DR, the vector of derivatives of the link function evaluated at $\hat{\pi}_i$.

3. Calculation of %VA, the vector of estimated variances of the binomial variables. Each component of the vector %VA is of the form $N_i \hat{\pi}_i (1 - \hat{\pi}_i)$.

4. Calculation of %DI, the log-likelihood or deviance for each observation. The sum of the %DI's for all the observations is a measure of goodness of fit, minimized by the maximum likelihood estimates. The formula for the deviance for all binomial links is the same as that for logistic regression (12, equation 6.16).

Although the first two calculations depend on the link function used, the last two are common to all links. Upon invocation of a single GLIM macro for a given link function, these calculations become transparent to the user.

GLIM MACROS

The GLIM macros RD, RR, HR, and OR for estimating the regression coefficients for identity, logarithmic, log-complement, and logarithmic links, respectively, are found in Appendix 1. Invocation of one of these macros by the \$USE macro command

TABLE 3
Comparison of fitted values* for different links

Cell	Observed	Parameter			
		OR	RR	RD	HR
1	11	10.6	10.4	11.7	11.8
2	5	5.9	5.9	5.6	5.6
3	11	10.5	10.6	10.0	10.0
4	6	5.7	5.7	5.4	5.4
5	3	1.6	1.6	1.6	1.6
6	1	2.7	2.7	2.9	2.9
7	4	3.7	3.7	3.6	3.6
8	3	2.5	2.5	2.4	2.4
9	12	13.8	13.7	13.8	13.8
10	4	4.5	4.7	3.7	3.7
11	2	1.2	1.2	1.1	1.1
12	6	5.3	5.3	5.3	5.3
13	0	2.4	2.4	2.5	2.5
14	1	1.8	1.8	1.9	1.9
15	12	7.8	7.7	8.7	8.7
16	7	5.1	5.3	4.4	4.4
17	2	3.0	3.0	2.9	2.9
18	8	9.9	9.8	10.5	10.5

* Fitted values are from model D which includes SMO, SOC, and ALC.

TABLE 4
Estimates of alcohol effect, comparing heavy drinkers with others, adjusted for smoking and social class

	Parameter			
	Log odds ratio	Log risk ratio	Risk difference	Health ratio
Estimate	0.7374	0.6830	0.0767	0.0868
Standard error	0.2432	0.2022	0.0351	0.0297

invokes other macros needed to perform the calculations for the given link needed to obtain the appropriate estimates for various models using \$FIT commands. (Although logistic regression can be done in GLIM more simply without macros (4), the OR macro is provided for comparison.) A given link remains in force until a new link macro or a \$ERR command is invoked.

Estimates of the regression parameters and their standard errors as well as the appropriate likelihood ratio tests are obtained in the usual manner. However, the usual confidence intervals based on the normal approximation and the reported standard errors can be misleading for asymmetric links such as the logarithmic. For

example, the parameter values within the confidence bounds may yield estimated probabilities which are not between 0 and 1. Details about this as well as how to use GLIM and the interpretation of the macro code can be found in the GLIM manual (4).

EXAMPLE

Wright et al. (13) investigated the relationship between maternal alcohol consumption and the risk of a low birth weight baby. Table 1 presents the total number of births and the number of births below the tenth percentile in birth weight in each of 18 categories defined according to whether or not the mother smoked, her membership in one of three social class groups, and by whether her drinking was light, moderate, or heavy. In their analysis, Wright et al. used the Mantel-Haenszel method (3) to account for the effects of social class and smoking. They noted that the estimates of the alcohol effect varied in the smoking and nonsmoking groups. English and Bower (14) suggested that logistic regression is the method of choice for this data set because it allows 1) tests for trend in ordered categories such as alcohol consumption, 2) simultaneous control of several confounders, and 3) tests of hypotheses of no interaction. Our methods retain these advantages, while allowing inference on parameters other than the odds ratio. Of particular interest is the assessment of interaction which depends on the link chosen. The GLIM commands used to obtain some of the results are found in Appendix 2. The annotated output is found in Appendix 3. Tables 2-4 compare the results from different links.

These macros were used to fit the low birth weight data. The unadjusted regression estimates for the logarithms of the risk ratios are 0.106 and 0.660 comparing the moderate and heavy drinkers, respectively, with the light drinkers. These are simply the logarithms of the crude risk ratios of 1.11 and 1.93. After simultaneous adjustment for smoking and social class category, the regression estimates increase slightly to

0.175 and 0.680. The goodness-of-fit test of the adequacy of the model ($\chi^2 = 13.61$, 12 df) indicated an acceptable fit. Treating alcohol category as a continuous variable with possible scores of 1, 2, or 3 while still adjusting for smoking and social class category suggests that the risk of giving birth to a low birth weight baby increases by a factor of 1.40 for each level. The implicit equal spacing between light and moderate and between moderate and heavy drinkers, however, does not seem justified because the risks to the babies of moderate and light drinking mothers seem very similar, while there seems to be a more substantial increase in risk to babies of heavy drinking mothers. The likelihood ratio tests of the alcohol effect are significant for all the models suggested. Wright et al. (13) suggest that the alcohol effect may depend on smoking category. As English and Bower (14) show for a logistic link, however, there is no significant improvement in the fit from including a smoking-alcohol interaction in any of the link functions considered.

The odds ratio estimates, tests, and fitted values are quite similar to those for the risk ratio, as would be expected for data such as these in which the probabilities are near 0.1 for all cells. There is a similar correspondence for the risk difference and health ratio since π is closely approximated by $-\log(1 - \pi)$ for small π . Although different interpretations of the parameter estimates for main effects and interactions would be given, the conclusions to be drawn from the various models considered in this example are not substantially different. There could be greater discrepancy in some of the significance levels if the sample sizes were smaller and in the fitted values from the various links if the range of the proportions over the 18 cells were wider.

CONCLUSIONS

The macros described above allow estimation of risk ratio and risk difference

parameters for binomial data while accounting for confounders and effect modifiers and retaining the simplicity, convenience, and flexibility of GLIM. For instance, continuous and categorical covariates can be included in the model, and subjects can be grouped according to covariate value, as in the example, or each subject can be retained as a separate unit. Estimates can be obtained for other link functions by simple extensions of this procedure.

REFERENCES

1. Rothman KJ, Boice JD. Epidemiologic analysis with a programmable calculator. Washington, DC: US GPO, 1979. (NIH publication no. 79-1649).
2. Breslow NE. Elementary methods of cohort analysis. *Int J Epidemiol* 1984;13:112-15.
3. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI* 1959;22:719-48.
4. Baker RJ, Nelder JA. The GLIM system release 3 manual. Oxford: Numerical Algorithms Group, 1978.
5. Weinberg CR. Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. *Am J Epidemiol* 1986;123:162-73.
6. Storer BE, Wacholder S, Breslow NE. Maximum likelihood fitting of general risk models to stratified data. *Appl Stat* 1983;32:172-81.
7. Thomas DC. General relative risk models for survival time and matched case-control analysis. *Biometrics* 1981;37:673-86.
8. Walter SD, Holford TR. Additive, multiplicative, and other models for disease risks. *Am J Epidemiol* 1978;108:341-6.
9. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980;112:467-70.
10. Siemiatycki J, Thomas D. Biological models and statistical interactions: an example from multi-stage carcinogenesis. *Int J Epidemiol* 1982;10:383-7.
11. McCullagh P, Nelder JA. Generalized linear models. London: Chapman and Hall, 1983.
12. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. Lyon: International Agency for Research on Cancer, 1980.
13. Wright JT, Barrison IG, Lewis IG, et al. Alcohol consumption, pregnancy and low birthweight. *Lancet* 1983;1:663-5.
14. English D, Bower C. Alcohol consumption, pregnancy, and low birthweight. *Lancet* 1983;2:1111.

Appendix 1
GLIM macros

\$SUBFILE MACROS \$

\$C N.B.: The binomial denominator is assumed to be N.

\$C**RD for RISK DIFFERENCE parameters*****

\$MAC RD \$PR ' ESTIMATION OF RISK DIFFERENCES' \$

\$OWN RDFV VA RDDR DI \$SCALE 1 \$CALCULATE %LP=.5 \$

\$C P=.5 is the starting value. \$ENDM \$

\$C**RR for RISK ratio parameters*****

\$MACRO RR \$PR ' ESTIMATION OF RISK RATIOS' \$

\$OWN RRFV VA RRDR DI \$SCALE 1 \$CALCULATE %LP=-.5 \$

\$C P=%EXP(-.5) is the starting value. Exponentiate regression coefficients to get the RISK RATIOS. !

\$ENDMACRO \$

\$C**HR for the HEALTH ratio*****

\$MACRO HR \$OWN HRFV VA HRDR DI \$SCALE 1 \$CALCULATE %LP=-.5 \$

\$C P=%EXP(-.5) is the starting value \$

\$ENDMACRO \$

\$C**OR for the ODDS RATIO*****

\$MACRO OR \$C 'ERROR BIN N' does the same thing as this macro \$

\$PRINT ' ESTIMATION OF THE ODDS RATIO' \$

\$OWN ORFV VA ORDR DI \$SCALE 1 \$CALCULATE %LP=0 \$ENDMACRO \$

\$C*****

\$MACRO RDFV \$C*****FITTED VALUES for RD**\$

\$C guard against probabilities which are out of range \$

\$CALCULATE %LP=%IF(%LT(%LP,0),.0001,%LP):

%LP=%IF(%GT(%LP,1),.9999,%LP):

P=%LP :%FV=N*P \$ENDMACRO\$

\$C*****

\$MACRO RRFV \$C*****FITTED VALUES for RR**

\$C Guard against probabilities below 0 \$

\$CALCULATE %LP=%IF(%GT(%LP,0),-.0001,%LP):

P=%EXP(%LP) :%FV=N*P \$ENDMACRO\$

\$C*****

\$MACRO HRFV \$C*****FITTED VALUES for HR** \$

\$C guard against probabilities below 0 \$

\$CALCULATE %LP=%IF(%GT(%LP,0),-.0001,%LP):

P=1-%EXP(%LP) :%FV=N*P \$ENDMACRO\$

\$C*****

\$MACRO ORFV \$C *****FITTED VALUES*for OR ** \$

\$CALCULATE ELP=%EXP(%LP):P=ELP/(1+ELP) :%FV=N*P \$ENDMACRO \$

\$C*****

\$C*****

\$MACRO VA \$C ***Variance of binomial proportion with N replications***\$

\$C applies to all links *****\$

\$CALCULATE %VA=%FV*(1-P) \$END \$

\$C*****

\$C*****

\$MACRO RDDR \$C **Derivative of the IDENTITIY function evaluated at %FV*!

\$CALCULATE %DR=1/N \$ENDMACRO \$

\$MACRO RRDR \$C ***Derivative of LOG function evaluated at %FV *****\$

\$C

\$CALCULATE %DR=1/%FV \$ENDMACRO\$

\$C*****

\$MACRO HRDR \$C ***Derivative of LOG function*** !

\$C

\$CALCULATE %DR=0.-1./(N-%FV) \$ENDMACRO\$

\$C *****

Continued

```

$MACRO ORDR $C ***Derivative of the LOGIT function evalauted at %FV** !
$CAL %DR=1/%VA $ENDMACRO$
$C*****
$C*****
$MACRO DI $C ***%DI is the DEVIANCE function (KU-KULLBACK INFO)**** $
$WARN $!
$CALCULATE %DI=-2*(%YV*%LOG(%FV/%YV)+(N-%YV)*%LOG((1-P)*N/(N-%YV))) !
$WARN $ !
$ENDMACRO $
$C*****
$C*****
$FINISH $

```

Appendix 2

GLIM Input Commands

```

$INPUT 1 72 $C READ MACROS FROM UNIT 1 (MUST BE DEFINED LOCALLY) $
$C Data from Wright et al. (13) $
$*****
$ECHO $OUTPUT 15 $C send output to unit 15 (locally defined)$
$UNIT 18 $C There are 18 proportions $
$DATA D N $C D is the no. of diseased babies out of N births. $
$READ 11 84 5 79 11 169 6 28 3 13 1 26
      4 22 3 25 12 162 4 17 2 7 6 38
      0 14 1 18 12 91 7 19 2 18 8 70 $
$CALCULATE ALC=4-%GL(3,1):SOC=%GL(3,6):SMO =%GL(2,3) $
$LOOK D N ALC SMO SOC $
$C E.g.: there were 11 low birthweight babies out of
      84 births to mothers who had heavy alcohol consumption,
      did not smoke and were in the lowest social class. $
$FACTOR ALC 3 SOC 3 SMO 2 $CALC ALC2=ALC $
$C ALC=alcohol consumption category, SMO=smoking category, and
      SOC=social class category are retained as categorical variables,
      or FACTORS. ALC2 is the continuous version of ALC. $
$YVAR D $C The DEPENDENT variable is D.$

$MAC STAR $PR '*****' $C separate fits $ENDMAC $

$USE RR $C to show the flexibility $
$FIT %GM $DISPLAY E $USE STAR $
$FIT ALC $DISPLAY E $USE STAR $
$FIT SMO+SOC $C deviance only $USE STAR$
$FIT +ALC $DISPLAY ER $USE STAR$
$FIT +ALC.SMO $DISPLAY E $USE STAR $
$FIT ALC2+SMO+SOC $DISPLAY E $USE STAR $
$C*****
$USE RD $C Invoke macros for fitting risk differences $
$FIT SOC+ALC+SMO $DISPLAY E $USE STAR $
$FIT +ALC.SMO $DISPLAY E $USE STAR $
$C*****
$ERROR B N $C Usual logistic regression $
$FIT SOC+ALC*SMO $DISPLAY E $USE STAR$
$STOP $

```

Appendix 3

Annotated GLIM output produced by code in Appendix 2*

```

$UNIT 18 $C There are 18 proportions $
$DATA D N $C D is the no. of diseased babies out of N births. $
$C Data from Wright et al. (13) $
$READ 11 84 5 79 11 169 6 28 3 13 1 26
      4 22 3 25 12 162 4 17 2 7 6 38
      0 14 1 18 12 91 7 19 2 18 8 70 $
$CALCULATE ALC=4-%GL(3,1):SOC=%GL(3,6):SMO =%GL(2,3) $
$LOOK D N ALC SMO SOC $
      1 11.00 84.00 3.000 1.000 1.000
      2 5.000 79.00 2.000 1.000 1.000
      3 11.00 169.0 1.000 1.000 1.000
      4 6.000 28.00 3.000 2.000 1.000
      5 3.000 13.00 2.000 2.000 1.000
      6 1.000 26.00 1.000 2.000 1.000
      7 4.000 22.00 3.000 1.000 2.000
      8 3.000 25.00 2.000 1.000 2.000
      9 12.00 162.0 1.000 1.000 2.000
     10 4.000 17.00 3.000 2.000 2.000
     11 2.000 7.000 2.000 2.000 2.000
     12 6.000 38.00 1.000 2.000 2.000
     13 .0 14.00 3.000 1.000 3.000
     14 1.000 18.00 2.000 1.000 3.000
     15 12.00 91.00 1.000 1.000 3.000
     16 7.000 19.00 3.000 2.000 3.000
     17 2.000 18.00 2.000 2.000 3.000
     18 8.000 70.00 1.000 2.000 3.000
$C E.g.: there were 11 low birthweight babies out of
      84 births to mothers who had heavy alcohol consumption,
      did not smoke and were in the lowest social class. $
$FACTOR ALC 3 SOC 3 SMO 2 $CALC ALC2=ALC $
$C ALC=alcohol consumption category, SMO=smoking category, and
      SOC=social class category are retained as categorical variables,
      or FACTORS. ALC2 is the continuous version of ALC. $
$YVAR D $C The DEPENDENT variable is D.$
$MAC STAR $PR '*****' $C separate fits $ENDMAC $
$C*****
$USE RR $C to show the flexibility $
      ESTIMATION OF RISK RATIOS
$FIT %GM $DISPLAY E $USE STAR $
      SCALED
      CYCLE DEVIANCE DF
      5 33.21 17
      ESTIMATE S.E. PARAMETER
      1 -2.217 .9531E-01 %GM
      SCALE PARAMETER TAKEN AS 1.000
*****
$FIT ALC $DISPLAY E $USE STAR $
      SCALED
      CYCLE DEVIANCE DF

```

Continued

```

3      23.96      15

      ESTIMATE      S.E.      PARAMETER      comments
1  -2.409      .1349      %GM
2   .1062      .2729      ALC(2)      exp(0.106)=1.11
3   .6596      .2092      ALC(3)      exp(0.660)=1.93
SCALE PARAMETER TAKEN AS      1.000
*****
$FIT SMO+SOC      $C deviance only $USE STAR $
      SCALED
CYCLE  DEVIANCE      DF
3      22.85      14
*****
$FIT +ALC      $DISPLAY  ER $
      SCALED
CYCLE  DEVIANCE      DF
3      13.61      12      The likelihood ratio statistic
                          is 22.85-13.61=9.24 (2 d.f.)
      ESTIMATE      S.E.      PARAMETER
1  -2.764      .2030      %GM
2   .5000      .2016      SMO(2)
3   .2926      .2331      SOC(2)
4   .2997      .2434      SOC(3)
5   .1749      .2741      ALC(2)      exp(.175)=1.19
6   .6802      .2154      ALC(3)      exp(.680)=1.97
SCALE PARAMETER TAKEN AS      1.000

UNIT  OBSERVED      FITTED      RESIDUAL
1     11.00      10.45      .1808      The 'Fitted' column
2     5.000      5.931      -.3976      from this table is the
3     11.00      10.65      .1100      RR column in Table 3.
4     6.000      5.744      .1196
5     3.000      1.609      1.171
6     1.000      2.702      -1.094
7     4.000      3.668      .1898
8     3.000      2.515      .3225
9     12.00      13.68      -.4752
10    4.000      4.673      -.3657
11    2.000      1.161      .8526
12    6.000      5.291      .3322
13     .0      2.351      -1.681
14    1.000      1.824      -.6435
15    12.00      7.741      1.600
16    7.000      5.260      .8919
17    2.000      3.007      -.6362
18    8.000      9.817      -.6253
*****
$FIT +ALC.SMO      $DISPLAY  E $USE STAR $
      SCALED
CYCLE  DEVIANCE      DF
2      12.02      10      Likelihood ratio (interaction)=
                          13.61-12.02=1.59
      ESTIMATE      S.E.      PARAMETER
1  -2.664      .2118      %GM
2   .2404      .2973      SMO(2)
3   .2662      .2324      SOC(2)
4   .2969      .2407      SOC(3)
5  -.4966E-01      .3611      ALC(2)
6   .4883      .2966      ALC(3)
7   .5733      .5486      SMO(2).ALC(2)
8   .4436      .4319      SMO(2).ALC(3)
SCALE PARAMETER TAKEN AS      1.000

```

Continued

```

*****
$FIT ALC2+SMO+SOC $DISPLAY E $USE STAR $
      SCALED
      CYCLE  DEVIANCE      DF
        2    14.03        13

      ESTIMATE      S.E.      PARAMETER
        1   -3.131      .2729      %GM
        2    .3353      .1088      ALC2      exp(.335)=1.40
        3    .5073      .2006      SMO(2)
        4    .3051      .2314      SOC(2)
        5    .2997      .2422      SOC(3)
      SCALE PARAMETER TAKEN AS      1.000
*****
$C*****
$USE RD $C Invoke macros for fitting risk differences $
  ESTIMATION OF RISK DIFFERENCES
  ----- CURRENT DISPLAY INHIBITED
$FIT SOC+ALC+SMO $DISPLAY E $USE STAR$
      SCALED
      CYCLE  DEVIANCE      DF
        3    14.92        12

      ESTIMATE      S.E.      PARAMETER
        1   .5885E-01   .1615E-01   %GM
        2   .2656E-01   .2323E-01   SOC(2)
        3   .3628E-01   .2700E-01   SOC(3)
        4   .1255E-01   .2572E-01   ALC(2)
        5   .8032E-01   .3028E-01   ALC(3)
        6   .5459E-01   .2705E-01   SMO(2)
      SCALE PARAMETER TAKEN AS      1.000
*****
$FIT +ALC.SMO      $DISPLAY E $USE STAR $
      SCALED
      CYCLE  DEVIANCE      DF
        2    11.64        10

      ESTIMATE      S.E.      PARAMETER
        1   .6374E-01   .1686E-01   %GM
        2   .2782E-01   .2328E-01   SOC(2)
        3   .3684E-01   .2690E-01   SOC(3)
        4   -.3640E-03   .2726E-01   ALC(2)
        5   .5466E-01   .3342E-01   ALC(3)
        6   .1907E-01   .3088E-01   SMO(2)
        7   .8379E-01   .7395E-01   ALC(2).SMO(2)
        8   .1082      .6963E-01   ALC(3).SMO(2)
      SCALE PARAMETER TAKEN AS      1.000
*****
$C*****
$ERROR B N $C Usual logistic regression $
  ----- CURRENT DISPLAY INHIBITED
$FIT SOC+ALC*SMO $DISPLAY E $USE STAR$
      SCALED
      CYCLE  DEVIANCE      DF
        3    11.97        10

      ESTIMATE      S.E.      PARAMETER
        1   -2.609      .2362      %GM
        2    .3137      .2686      SOC(2)
        3    .3481      .2816      SOC(3)
        4   -.4743E-01   .3935      ALC(2)
        5    .5565      .3361      ALC(3)
        6    .2606      .3352      SMO(2)
        7    .6771      .6348      ALC(2).SMO(2)
        8    .5821      .5152      ALC(3).SMO(2)
      SCALE PARAMETER TAKEN AS      1.000
*****
$C*****
$STOP $

```

*Upper case was produced by GLIM, lower case is annotation and comments.